ANOVA

Zack Kertcher

Statistics for Management
Fall 2016

Plan for today

1. Hypothesis testing (cont.)
2. Chi-square test
3. ANOVA
4. * Projects

Hypothesis testing terminology

Term and Use	Example
One-sample test Compare a single sample to a parameter	Do our machines perform as well as they should? Compare a random sample of product to an expected mean defects, per manufacturer's specification.
Two-sample test Comparing a parameter in two independent samples	Do our current machines perform as well as the new machines? Compare two random samples for mean defects.
Paired sample Comparing a parameter in non-independent samples	Do our machines perform as well after a year? Compare two samples from the same machines at the beginning and end of the year.
Type I error Cry wolf, when there is none.	You claim that there is a difference in the mean defects of your machines, when there isn't.
Type II error Don't cry wolf, when there is.	You claim that there is no difference in the mean defects of your machines, when there is.

Hypothesis testing with skewed/unknown distributions

- ✔ We already discussed repeated samples, but this is not always possible.
  (We might not know the population. It could be costly)
- ✔ Instead, we can use a nonparametric hypothesis test.
  (The Wilcoxon-Mann–Whitney U-test)
- ✔ A parametric hypothesis test is more "conservative"
  (Less sensitive to outliers, and can be applied to various distributions)

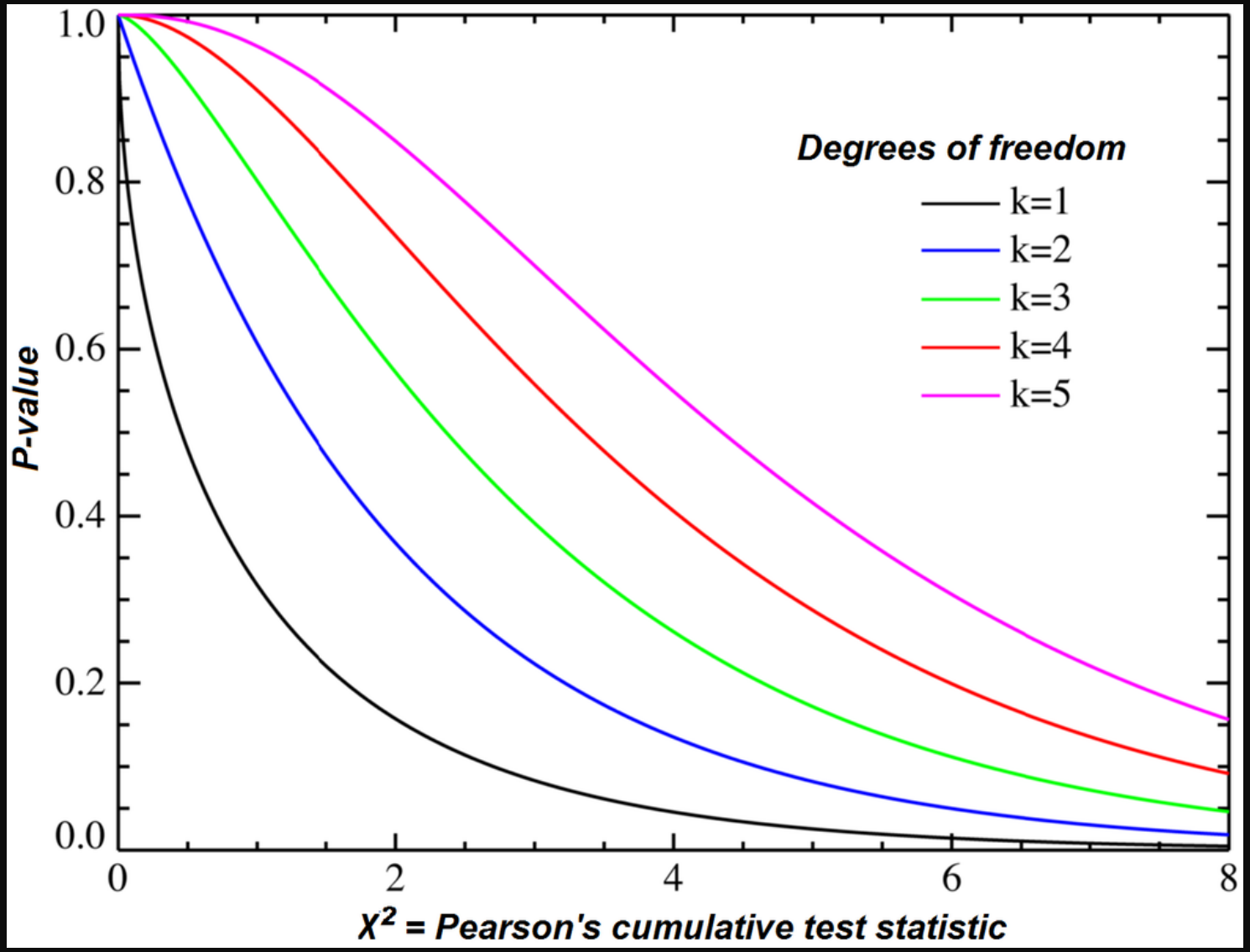
Non-parametric hypothesis testing example

CHI-SQUARE TEST

When to use it?

- ✔ When testing hypothesess about two factors (categorical variables)
- ✔ The null hypothesis is that the variables are independent
- ✔ Produces a Chi-squared statistics (χ^2)
  (Follows a Chi-squared distribution, with a degrees of freedom parameter)

Chi-squared distribution



Source: https://en.wikipedia.org/wiki/Chi-squared_distribution

Assumptions

- ✔ Observations are independent
- ✔ Sample size (n) is sufficiently large
  (When less than 20% of cells have a value < 5)

One-factor Chi-squared test example

You lead a team of 8 analysts. Reviewing your team's prediction performance (0-100) you see following data:

A1=98.4; A2=72; A3=96; A4=81; A5=77.5; A6=62; A7=69.5; A8=74.5

- ✔ How confident are you that your team's predication capability is equal, at 99% confidence?
- ✔ Would your conclusion change at 95% confidence?
- ✔ Would it change if you remove the lowest performing analyst?

Two factors Chi-squared test example

You own an HVAC company. You plan to enter a new market, and you want to know on which market segment to focus your advertising. You start by considering if there is a difference in the proportion of air conditioning installations in a sample of houses in desirable and less desirable locations.

- ✔ Load the **real estate** data as realestate
- ✔ Compare the presence of air conditioners (airco) in more and less desirable areas (prefarea)
- ✔ Save the output as a high-quality table with chi-squared statistics
- ✔ What is your conclusion?

ANOVA

When to use it?

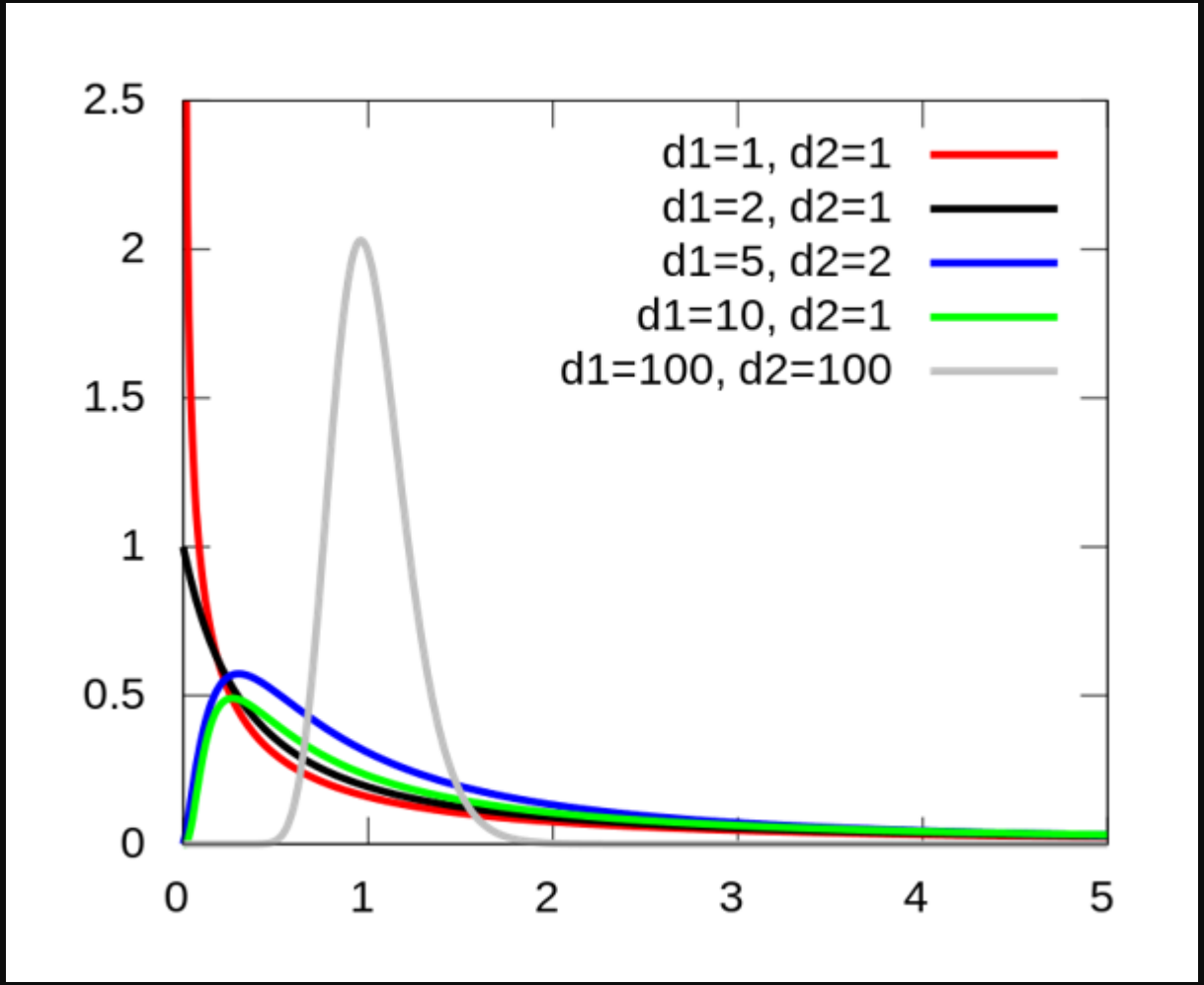
- ✔ Test hypotheses (e.g., compare means) related to more than two groups and a numeric DV
- ✔ Produces an F-statistic (which like Chi-square, has degrees of freedom, but uses two d.f., and follows an F-distribution)

F = mean square among / mean square within

variability between groups
(treatment effect)

variability within groups
(experimental error)

F distribution



Source: https://en.wikipedia.org/wiki/F-distribution

Assumptions

- ✔ Samples are independent
  (Not paired)
- ✔ Sample sizes are adequate
  (e.g., <10% of population and (n>10)
- ✔ Population distribution is approximately normal
  (This is critical when sample sizes are small)
- ✔ Variance in groups is about equal

One-way ANOVA example

You manage a sales team. You are doing an experiment that randomly splits the team into 6 equal groups of 10. Each group is asked to make a different number of calls to prospective clients (from 1 to 6 calls).

- ✔ Load the sales.csv data as **sales**
- ✔ What is your null hypothesis?
- ✔ Run an ANOVA test at 95% confidence level
- ✔ Check for normal distribution
- ✔ What is your conclusion about the experiment?

Additional analysis of variance-type tests

- ✔ If you consider more than a single factor (group) perform a two/three/N-way ANOVA test (MANOVA)
- ✔ When the data are skewed and you are testing a single factor use the non-parametric Kruskal–Wallis test (kruskal.test)
- ✔ When the data are skewed and you are testing multiple factors use the non-parametric Friedman test (friedman.test)